# 

**Logistic Regression - Stata**

Today we’re talking about how to run and interpret logistic regression models.

# 

# Logistic Regression

## Basics

We use logistic regression when we have a binary, dependent variable (outcome variable) and multiple independent variables. This is one of the most commonly used models in epidemiology! For example, you might want to understand the relationship between developing heart disease, and risk factors such as blood pressure, age, and height.

The logistic function is a probability, from 0 to 1 and is non-linear in its coefficients. We use the logit transformation to make the results easier to interpret.

### Measures of Association

There are two main measures of association for logistic regression:

1. Odds ratio (OR): ratio of the odds of the outcome when a risk factor is present, divided by the odds of the outcome when the risk factor is absent
2. Adjusted odds ratio: odds ratio between the risk factor and the outcome, after adjusting for the effect of the other independent variables in the model.

### Hypothesis testing for model parameters

There are two main hypothesis tests we can use in logistic regression:

1. Wald test statistic: tests for whether a particular regression coefficient equals zero
2. Likelihood ratio: tests for whether a set of coefficients are all simultaneously equal to zero

When you are testing for a single risk factor, the likelihood ratio test has better statistical properties than the Wald’s test, especially when the sample size is smaller.

## Two by Two Tables

Two by two tables are helpful for us to understand the relationship between an exposure and an outcome. For these tables, we can use the chi-square test to determine if there is a significant relationship, but this does not provide a measure of its relative magnitude. An odds ratio, however, does tell us about this magnitude.

### In Stata

The basic syntax for the logistic regression model is:

logistic *outcomevar exposurevar*

Where outcomevar is the health outcome of interest (dependent variable), and exposurevar is the risk factor of interest (independent variable). In order to get the coefficients for your model, use:

logit

*Note:* the Stata command logistic estimates the odds ratio, while logit gives the regression coefficients which you can use to write out your model equations.

## Interpretation of the OR

Interpretation of the OR is in terms of the odds of exposure, so the wording for the OR can be a little tricky. Here are a few examples of what interpretation of the OR can look like. Say we have an OR of 5 for heart disease.

* Cases have 5 times the odds of heart disease compared to controls.
* After adjusting for age, cases have 4 times the odds of having heart disease compared to controls.
* Cases that are a year older have 1.05 times the odds of disease than those that are a year younger

In your output, you will also receive the p-value for the likelihood ratio test in the “Prob > chi2 =” line in the upper right. This tells you whether there is a significant relationship between your exposure and outcome.

### Lincom in Stata

After running your model, you might want to understand what the odds ratio is for a 10-year increase in age, or a 5-unit increase in another variable. This is where the lincom statement helps.

1. First, run your logistic statement, then
2. Lincom (unit of interest)\*(variable)

For example, for a 10-year increase in age you would type: lincom 10\*age

This tells you: for a 10 year increase in age, cases that are 10 years older have 3.5 times the odds of heart disease than those that are 20 years younger.

### Multiple logistic regression interpretation

Most logistic regression models have multiple variables. To understand the ORs for each variable, you’ll want to look at the OR for each variable line. The p-value on each line of the output tells you if that specific variable has a significant association with the outcome. You can then use the overall likelihood ratio test to see if all of the variables have a significant association, just like before.

If they differ, say one variable is significant and another is not. The likelihood ratio test will also likely not be significant. This means that just that one variable is significantly associated with your outcome,

## Interaction and Confounding

In order to test whether there is interaction or confounding by another third variable, use the following steps.

1. Generate an interaction term between your exposure and third variable
   1. Generate inter=exposure\*thirdvariable
2. Run your logistic regression for your full model with the interaction term
3. Store those estimates
   1. Estimates store full
4. Run your logistic regression for the reduced model without the interaction term
5. Store those estimates
   1. Estimates store reduced
6. Run a likelihood ratio test of the estimate
   1. Lrtest reduced ful

If the p-value for the test is <0.05, you have interaction and should report the strata specific ORs. If it is pretty close to 0.05 say 0.07, you should still probably report them separately.

If you don’t have interaction, assess for confounding with these steps.

1. Identify the crude OR (the OR without adjustment of any variables) and the adjusted OR (the OR that includes adjustment for other variables)
2. Use the following equation to determine if there is confounding using the 10% rule. If the number is greater than 10, you have confounding.
   1. Display ((adjusted-crude)/adjusted)\*100